Statistics Canada | Statistique Canada

Health Division | Division de la santé

Statistics Canada

Health Division

Quality assessment of vital statistics

(A pilot study)

by

D.N. Nagnur
S.G. Currie
B. Heath

Published by authority of
The Minister of Supply and
Services Canada

Statistics Canada should be credited
when reproducing or quoting any part of
this document

© Minister of Supply
and Services Canada 1981

August 1981
4-2302-534

Ottawa

Version française de cette publication
disponible sur demande.

CONTENT ANALYSIS AND QUALITY ASSESSMENT OF VITAL STATISTICS DATA - A PILOT STUDY

Though the quality of the vital statistics data published by Statistics Canada has been assumed historically to be of high order, no attempt was made to assess quantitatively the quality of the data prior to this study. This assessment project was initiated under the rubric of "Content Evaluation and Quality Assessment of Integrated Vital Statistics (IVS)". The study included the events of births, stillbirths, marriages and deaths for the data year 1976.

Since each of the vital events has legal implications, the completeness of registration of these events has been justifiably assumed to be of high order.

The basic vital statistics data are compiled by the provinces and territories by means of registration forms designed for administrative and legal purposes. Under the federal-provincial agreements dating back to 1919, Statistics Canada has been employing these records to produce national data on mortality, natality, nuptiality and longevity.

Increasingly, government departments, private agencies, and public at large are employing statistics relating to vital events as important inputs in their planning and policy decisions, especially those related to health and social programme areas. Consequently the analysis and measurement of demographic and epidemiological trends reflected by the time series data on vital events have acquired added importance and immediacy.

In order to assess the quality of the vital data, several dimensions of the study may be visualized. However, the one followed was the sequence of processes in the generation of informational output from the initial compilation of data. Some of the important phases that also need to be studied are:

(a) timeliness of registration and the quality of the input from the individual registrants;

(b) the initial editing and the flow of registration forms from the network of regional centres to the provincial and territorial offices;

(c) the lag-time associated with the process of transmission of data and the intervening processes involved in translating the data into machine readable form;

(d) the modes of transforming data to the Vital Statistics and Disease Registries Section, Statistics Canada, and the effect it would have on the timeliness, quality etc. ...;

(e) the frequency and timeliness in incorporating or associating amendments to the original documents at the provincial level;

(f) after the receipt of data at the vital statistics offices, Statistics Canada, the data are then subject to several processes such as editing, imputation and ultimately the production of a multiplicity of tabulations. Computer routines are employed in some of the processes while manual manipulations are resorted to in the others.

In order to obtain a complete profile of quality of the vital statistics information, it is essential to analyse each of the sequential transition stages in the translation of "data" into "information". However, due to several limitations, the present exercise is confined to a small subset of the total enquiry. It begins with the phase wherein the data are received at the Vital Statistics and Disease Registries Section, Statistics Canada, from the provincial and territorial offices of vital statistics. It ends with the edited final computerized record system employed for generating tabulations and reports.

The sampling procedure that was developed and adopted, involved stratification primarily based on the mode of data availability from the provinces for further processing and tabulation. The sample included 4,400 records distributed among all the provinces - 1,300 for each of the events of birth, marriage and death and 500 with respect to stillbirths. These samples constituted 0.2% of the annual events for births, 0.7% for marriages, 0.8% for deaths and about 19% for stillbirths. The aggregate sample for all events combined was about 0.6% of the total events for 1976. The detail examination was confined to the key fields representing what were designated as core characteristics in each series.

The core characteristics studied with respect to different events were:

Births

Sex of the child
Date of birth
Kind of birth
Order of birth

Residence of mother
Duration of pregnancy
Total children born
Children liveborn
Children stillborn
Birthweight of child
Parents married to each other
Marital status of mother
Birthdate of mother
Age of mother

Stillbirths

Sex of the child
Date of stillbirth
Kind of birth
Order of birth
Residence of mother
Duration of pregnancy
Number of children born
Number liveborn
Number stillborn
Birthweight
Marital status of mother
Parents married to each other
Birthplace of mother
Birthplace of father
Birthdate of mother
Age of mother
Cause of stillbirth

Marriages

Date of marriage
Marital status of groom
Birthdate of groom
Age of groom
Marital status of bride
Birthdate of bride
Age of bride

Deaths

Sex
Place of death
Residence of deceased
Marital status
Birthdate
Age
Date of death
Cause of death
Autopsy
Date of certifications

Initially, it was proposed to include in the study a three-way match. The records from initial data received from the provinces were to be matched with those recoded from the microfilm copies of original documents and the final edited tape. Because of some unavoidable difficulties, this study was confined to only two-way match - the records from the final edited tape were matched with the recoded records from the original microfilm documents. It is planned to adhere to the original proposal in the follow-up studies for the later years.

Results

The results indicate that an acceptable level of accuracy and reliability are being preserved and maintained through the various phases of internal processing at the Vital Statistics and Disease Registries Section, Statistics Canada. The error rate measured on the basis of adjudicated "mismatches" with respect to the majority of core characteristics indicate a reasonable degree of accuracy and consistency.

The report prepared by S.G. Currie of the methodology division provides the detail breakdown with respect to various characteristics. The report is attached in the Appendix 1. A brief summary of the results for each series is provided here.

Marriages

The aggregate error rate for the seven core characteristics for marriage was 1.33 per 100 documents with none of the individual characteristics registering an error rate in excess of 0.5 per 100 documents. The fields that gave relatively higher error rate were those representing birthdate, marital status and age of groom; the latter two were of small order of magnitude, about 0.25 per 100 documents.

Also there was no substantial difference between the provinces reporting on tape and those reporting on microfilm or forms. It was also found that a sizable portion of the discrepancies was due to missing information and eventual imputation.

Births

The aggregate error rate for all the core characteristics for birth data, excluding the fields on the number of children born, was 2.7 per 100 records. The highest individual field error rate was about 1% for "residence of the mother". The second highest rate was about 0.9 per 100 records for "birthweight" of the child. These two fields also experienced non-response rates of 0.5% and 1.5% respectively. These characteristics need to be further investigated. With the expected reduction in the frequency of non-response, it is likely that error rate will diminish further.

Other fields that experienced significant discrepancy were those representing "total number of children", number of children stillborn and to a lesser extent the number of children liveborn. The discrepancies, however, do not necessarily translate to the same extent into error components in the tabulated data. Some discrepancies appear to arise as a result of the "wording" of the questions eliciting information on the number of children. The wording differs among provinces. Part of the problem may also be attributed to missing data and subsequent imputation. These aspects need further investigation and will be studied in follow-up rounds.

The provinces reporting on microfilms or forms only experienced slightly better error rates than those reporting on tapes. In order to isolate errors further, coding and other aspects, especially with reference to fields representing "residence of the mother", "birthweight" of the child and "duration of pregnancy" will be carefully monitored in the next round of enquiry.

Deaths

The results with respect to "deaths" should be read with caution, because, though some of the error rates taken individually and in aggregate appear appreciable, the effects these would have on the important tabulations are not in corresponding proportions.

For example, the error rate of more than three per 100 documents with respect to autopsy does not have any effect on any of the mortality tabulations that are normally produced. Similarly, place of death which accounts for an error rate in excess of five per 100 records does not have any bearing on the ultimate tabulations since all the "death" tabulations are on the basis of "normal place of residence". The high error rate with respect to "residence" of the deceased would have an effect but the detail analysis reveals that it is mainly at the "locality level," and at the levels of aggregation to County, Census Division or higher order, the error diminishes significantly. Since most of the mortality tabulations are at the national or provincial level, the error rate as revealed by the study with respect to "residence" is less serious than it actually appears. The error rate with respect to underlying cause of death is some cause for concern; hence, it seems necessary to carefully monitor the sources of this error so that these can be minimized and rectified. It seems all the more important since this classification forms the basis for many epidemiological and medical follow-up exercises. With this aspect in mind, the follow-up round of enquiry is focussed on monitoring the cause of death alone for the 1979 death data. The results of that study are expected to throw further light on the error patterns specific to the underlying cause of death.

The overall error results with respect to deaths may be considered as slightly on the higher side relative to "marriages" and "births"; but, this may be expected as the personal data regarding the deceased has to be necessarily provided by the relatives and others and modifications to the initial data or amendments are obviously more frequent. Imputation of certain fields like "birthdate" also adds to the aggregate error, though, it does not have any significant effect when the age is independently reported, for the purposes of tabulations ... several aspects of death data will be monitored on the successive rounds of enquiry.

Stillbirths

Though the completeness of coverage of stillbirths is of significant importance, the detail characteristics associated with them may be less so in comparision with births, marriages and deaths. The most important indicator in which the number of stillbirths is an important component is that of perinatal mortality. However, other characteristics with respect to stillbirths are of significant importance in studying the pattern of late foetal deaths, their causes and consequences.

The errors observed in "stillbirth data" were found to be concentrated in the fields relating to the number of children and the underlying cause of death. The number of children born affects the tabulations of births by total birth order and one is inclined to suggest that when considering the distribution of births by birth order, the one with respect to live birth order is to be preferred to the one by total birth order. Efforts should also be directed in improving the statistics on the number of children born-live as well as stillborn.

Another significant field with appreciable level of error is that representing cause of stillbirth. A closer examination reveals that the error rate was less severe than is shown by the results, since more than half of those cases contributing to the error, agreed in the first two digits of the 4-digit categories. Residence of the mother as in respect to other events, differed more significantly at the locality level and at an aggregate level of county or census division and higher order, this error did not constitute a very significant factor.

The General Conclusions

The broad conclusion that emerges from the study is that the impact of errors associated

with internal processing is not generally serious though there are areas and characteristics where improvements are necessary and advisable and would undoubtedly result in the derivation and dissemination of better quality information. However, there do exist some marked regional differences in the quality of data with respect to certain characteristics and certain events; one of the contributing factors is the missing information or non-response items. These could be corrected and improved possibly by close monitoring of the data at collection or compilation stage and by strengthening a system of follow-up enquiry. The study also points up the fact that detailed investigation, with initially proposed three-way match, and with increased sample size to provide reliable measurements at the provincial level should be carried out. In order to obtain a complete profile of the quality of the vital statistics information, the analysis of each of the sequential transition stages in the process of translation of "data" into "information", as suggested earlier, should be undertaken. The quality assessment study should logically lead to setting up of a mechanism of quality control and feed back procedures, on current data year basis, to the provincial respondent agencies. Once the benchmark studies to establish the pattern of reliability measures, nationally as well as provincially, for different vital events distributed by core characteristics are completed, then, follow-up studies could be oriented to focus on the specialized aspects of vital statistics data system.

As a complement to this current quality assessment exercise, statistical formulae for the measurement of variance and standard errors for the most commonly employed mortality and fertility rates were investigated. A technical note on this may be obtained on request. Beginning probably data year 1979, some of the variance and standard error measurements will be incorporated as part of our effort to improve the contents and quality of the annual reports on births and deaths.

APPENDIX I

EVALUATION OF VITAL STATISTICS PROCESSING

MARRIAGES, BIRTHS, STILLBIRTHS AND DEATHS

The quality of vital statistics data has always been considered good although no direct measurements had been made prior to the current evaluation study. The measurement of data quality by means of such studies is in accordance with the recently introduced bureau policy on data quality. To the extent that the results of this and future evaluation provide meaningful measures of data quality, they will be incorporated into vital statistics publications. The results may also provide a basis for methodological improvements in data collection and processing.

Specifically the current study was designed to measure the levels of error introduced by coding, keying and edit correction operations during the processing of vital statistics records. Since much of the coding and keying of the data is now carried out by the provinces, the need for a monitoring system to ensure that quality standards are met has become a pressing concern for Statistics Canada. The current study should not be regarded as a one-time effort to measure data quality but rather it should be seen as the starting point for developing a continuing system for monitoring the quality of vital statistics data.

Methodology

The measurement of processing error rates was based on a sample of records from each series which were independently reprocessed and compared with the original production records on the clean data base. Only key fields were examined.

It was expected that error rates might vary from province to province since the provinces do varying amounts of processing using slightly different forms and procedures. Samples were selected as follows:

	Births	Marriages	Stillbirths	Deaths
Provinces reporting on tape:				
Quebec	326	329	-	334
Other	324	322	287	329
Provinces reporting on microfilm or forms:				
Newfoundland	319	305	74	300
Other	313	318	151	322
Total	**1,282**	**1,274**	**512**	**1,285**

Quebec and Newfoundland were sampled separately as it was felt that these two provinces might have peculiar problems. Samples from the two groups of provinces were allocated to each province proportionally.

The samples were selected using the registration numbers rather than actually physically selecting from microfilm or computer records. The range of registration numbers was determined for each province and this information along with the desired sample sizes was input to a sample selection program which produced a list of sample registration numbers. The records corresponding to selected registration numbers were located on the microfilms and photocopies produced.

Since the original coding of records appeared on the microfilm it was necessary to blank out the codes on the photocopies in order to make the recoding independent of the original. The independence of the recoding operation is extremely important in achieving accurate error rates since it is known that recoders will tend to agree with the original code more often if they know what it was. The recoders recorded their entries on specially designed coding sheets which were then keyed. The keying was 100% verified to ensure that keying errors were minimized.

The original methodology had specified that the recoded records be matched against the raw data tapes as well as the cleaned edited file. Unfortunately the raw data tapes were not easily accessible for this purpose and had to be dropped from the matching routine. This was unfortunate as it made the identification of types of errors much more difficult.

A match between the recoded records and the corresponding records on the clean edited file was carried out and a listing of all differences produced. An adjudicator reviewed the differences and assigned errors. Unfortunately the adjudicator did not have access to edit correction information as this proved to be too time-consuming to retrieve. Hence judgments were based on the microfilm records and a list of amendments:

The results of this study were weighted in order to combine estimates from the strata which were sampled. This was necessary as the sampling fractions varied considerably. The weight for the i^{th} stratum was computed as

$$Wi = \frac{Ni}{\sum_{1} Ni}$$

where Ni is the total number of registrations in the i^{th} stratum. The combined estimate of an error rate is given by

$$ec = \frac{\sum Wi.ei}{\sum_{1} Wi}$$

where ei is the error rate in the sample from the i^{th} stratum.

Note that $\sum_{1} Wi=1$ for national estimates but would be less than one when estimating an error rate for provinces reporting on tapes. The values of Ni used in computing weights in this report were preliminary and will not agree exactly with weights computed from published totals.

Results for Marriages

The results were very good for all the individual fields which were checked. However, some deficiencies in the data are not reflected in the sample results. Both British Columbia and Ontario do not report birthdates. Since the birthdate fields have higher error rates than other key fields, these provinces appear better than they really are. As well, British Columbia does not report the day of marriage on its tapes (only month and year) although the day is given on the microfilm.

It was not possible to identify keying errors in the original processing but the results of the re-processing suggest that keying is probably not a serious problem. The only keying error identified was due to a single marriage record in which the data had been shifted one position causing errors in three fields.

Another problem which is not reflected in the sample results is the method of imputing age from birthdate. The computer imputation routine calculates the age using the year of birth and the year of marriage rather than the complete dates. This resulted in many discrepancies in the sample which were not counted as errors. For example, in Manitoba, discrepancies occurred for about half the sample records.

Total non-response occurred when both date of birth and age were blank. Note that if birthdate is reported but age is blank, age is imputed from the birthdate in the edit program. Non-response on the microfilm occurred mainly for marital status. In these cases blank marital status was imputed to single if there was no indication of a previous marriage.

In comparing provinces reporting on tapes with those reporting on microfilm/forms, there is no evidence of any substantial differences in processing error rates.

Births

The dominant factor in the birth results is the high error rate for the three fields on "number of children born to this mother". Including these fields the error rate for all fields was 45.55 errors per 100 documents. But excluding them the error rate drops to a very respectable 2.70 errors per 100 documents.

Most of the errors on the three fields on "total children born to this mother" were not really coding errors in the strictest sense but differences between a code and a blank entry. The issue essentially revolves around how these fields should be treated when they are originally left blank. It is possible to provide entries either by correction or by imputation. It is likely that imputing a value was favoured over collecting more information by follow-up. The common imputations are as follows:

(1) If number of stillborn and total children born are blank, impute number stillborn = 0 and total children born = number liveborn.

(2) If number liveborn is blank, impute number liveborn = 1.

It is quite possible that such imputations introduce some bias but no firm conclusions can be drawn from the current study.

For the remaining fields, there was a high rate of discrepancies in which the code was assigned using information which was not on the original microfilm. These discrepancies occurred mainly for two fields, "duration of pregnancy" and "weight of child". For Ontario and British Columbia, "duration of pregnancy" was changed in a relatively consistent manner - low values were adjusted upward and high values were frequently changed. Whether additional information is obtained in these cases or whether values were amended in some predetermined manner needs to be further examined.

The field "weight of child" was a source of many discrepancies for Ontario. This occurred because the province always rounded up fractional ounces whereas the recoders rounded up or down in the usual way. This implies a small upward bias for "weight of child" which is probably not significant.

The fields "duration of pregnancy" and "weight of child" also experienced some non-response as 0.38% and 0.15% were missing from the microfilm but coded on the edited file while for "weight of child" there was no response at all in 1.52% of the cases.

The most error-prone field, excluding the fields on "number of children born", was "residence of mother" which had a processing error rate of 1.03%. In addition there was a 6.85% rate of other discrepancies, most of which arose because of poor reporting, including illegible and misleading responses as well as some non-response. There were also a few problems due to the use of a coding manual which needed updating.

Comparison suggests that provinces reporting on microfilm/forms experienced slightly better error rates than provinces reporting on tapes.

Results for Stillbirths

For stillbirths, the processing error rate per 100 documents was 13.10 for the three fields on "number of children born", 13.75 for "cause of death", 4.29 for "residence of mother" and 3.16 for other fields.

While the error rates for "number of children born" are better than those for births, this may be more a reflection of variability of the adjudication process than an indication of a real difference. There is no doubt that these data items are ambiguous especially in light of the fact the wording on some forms was changed with respect to whether or not the event being reported should be included. The reporting of multiple births was a source of confusion which can only be overcome by providing clearer instructions to persons completing the forms.

The error rate of 4.29% for "residence of mother" was higher than for births but still acceptable. It should be noted, however, that response errors accounted for discrepancies at a rate of 7.34%. This probably reflects less than acccurate reporting of municipality and to some extent codebook problems.

Non-response occurred for two fields. "Sex" had a blank rate of 1.24% on the microfilm; it appears that entries on the edited file were imputed manually during edit corrections. "Weight at birth" was blank on both the microfilm and the edited file 2.65% of the time. The lack of a birth weight for a stillbirth is significant in the sense that the weight, as well as the duration of the pregnancy, are used to determine the nature of the event, i.e., to distinguish a stillbirth from a miscarriage/abortion.

It should be noted that, although there was a low error rate for "weight at birth", the amendment rate of 6.72% was substantial. Almost all the amendments came in the Ontario data.

The two fields "parents married to each other" and "marital status of mother" gave rise to high imputation rates of 13.11% and 41.69% respectively. The system imputes "yes" for "parents married to each other" if "marital status mother" equals "married", and vice versa. The imputation of "parents married to each other" in this way is somewhat suspect as the mother may in certain cases be married to someone other than the father.

The field "birthdate mother" was imputed a phenomenal 56.09% of the time. In these cases the computer system created a "year of birth" rather than a "birthdate" using the age of the mother. The utility of this procedure and its accuracy need to be examined. The coding error rate for the underlying cause of death was rather high at 13.75%, substantially higher than the error rate for the death series. This perhaps is to be expected as the cause of death for stillbirths may be more poorly defined than in other deaths.

An analysis shows that in 31.9% of the errors there was agreement except on the fourth and last digit of the code and in another 24.6% the first two digits agreed. This suggests that the error rate is related to the level of detail of the coding.

Results for Deaths

The results for deaths are generally not as good as the other series. Clearly the geographic coding (place of death and residence deceased) was the major problem. The error rate of 9.00% for "residence deceased" is higher than the error rates for "residence mother" in the birth and stillbirth records. The error rate of 5.61% for "place of death" is also high for such a key data item. It appears that the poor geographic coding is related to poor coding manuals and imprecise reporting of place names.

The fields "birthdate" and "autopsy" with error rates of 2.89% and 3.25% respectively are worth noting. They have relatively high error rates for simple fields. It is not known how much of the error rates are due to coding and how much to keying but it is clear that similar fields on other types of records were processed with lower error rates.

Imputations affected two fields. The 11.48% imputation rate for "birthdate" reflects calculation of birthdate from age when the birthdate was not reported. For "date of death" the imputation rate was 10.56%. This resulted because of the non-reporting of "day of death" in some provinces which is imputed as "01" by the computer system.

The coding error rate of 7.18% for "cause of death" seems fairly reasonable, although it may be possible to do better with tight controls on this coding operation. About two thirds of the errors involved the first or second digit of the 4-digit code. This indicates that the errors were more serious than those encountered in the stillbirth coding.

Conclusions and Recommendations

On the basis of this study it appears that processing error rates vary from the extremely low rates for marriages to the moderately high rates for some fields in the stillbirth and death series, while the error rates are probably not high enough to be a serious problem, they are significant enough to justify further monitoring. On the other hand, there is ample evidence to suggest that item non-response and the methods of handling it are a more critical problem. It appears that the data are subject to varying degrees of amendment, correction and imputation, and that there is no assurance that some of these procedures are not introducing biases into the data.

The current study is deficient to the extent that it has not been able to precisely identify the sources of error. Had the original keyed data and the edit correction information been available, it would have been possible to make a much better assessment of the nature of the processing errors found. Given the lack of information upon which to base decisions, the accuracy and consistency of the adjudication process may have been a problem.

The following recommendations are offered:

(1) that the questions on "number of children born to this mother" be reviewed in order to obtain clear and consistent responses, especially for multiple births. If this cannot be done, then "birth order" statistics are seriously affected;

(2) that the procedures for doing geographic coding be reviewed and steps taken to ensure that provincial coders have the correct coding manuals and these are updated promptly;

(3) that the quality of the "cause of death" coding be monitored continuously using small samples which can be recoded by expert coders;

(4) that the nature of amendments to the data such as those for "duration of pregnancy" be investigated;

(5) that all adjustment and imputation routines in the processing be reviewed to ensure that they do not introduce serious biases;

(6) that causes of item non-response be identified and controlled. Where certain provinces do not collect certain key data items, there should be negotiations undertaken to include them on the forms in future years;

(7) that a study similar to the current one be carried out in the future using a methodology which will better identify the source of errors.

TEXT TABLE I. Marriages - Weighted Results

Field	Processing error	Other discrepancies(1)				
		CP	R	NR^1	NR^3	Total
	per cent					
Provinces reporting on tape:						
Date of marriage	.09	.09	.17	-	-	.26
Marital status - Groom	.26	-	-	.79	-	.79
Birthdate - Groom	.44	-	.44	.09	-	.53
Age - Groom	.26	-	.09	.09	.17	.35
Marital status - Bride	.09	-	-	.35	-	.35
Birthdate - Bride	.09	-	.35	.09	-	.44
Age - Bride	.09	-	.22	.09	.17	.48
Total	**1.32**	**.09**	**1.27**	**1.50**	**.34**	**3.20**
Provinces reporting on microfilm/forms:						
Date of marriage	.24	-	-	-	-	-
Marital status - Groom	.07	-	-	-	-	-
Birthdate - Groom	.73	-	-	-	-	-
Age - Groom	.24	.24	-	-	-	.24
Marital status - Bride	.24	-	-	-	-	-
Birthdate - Bride	-	-	-	-	-	-
Age - Bride	-	-	-	-	-	-
Total	**1.52**	**.24**	**-**	**-**	**-**	**.24**
All provinces:						
Date of marriage	.10	.08	.16	-	-	.24
Marital status - Groom	.24	-	-	.71	-	.71
Birthdate - Groom	.46	-	.40	.08	-	.48
Age - Groom	.26	.02	.08	.08	.16	.34
Marital status - Bride	.10	-	-	.32	-	.32
Birthdate - Bride	.08	-	.32	.08	-	.40
Age - Bride	.08	-	.20	.08	.16	.44
Total	**1.32**	**.10**	**1.16**	**1.35**	**.32**	**2.93**

(1) CP - Coded by province from other information.
R - Reporting problem.
NR^1 - Non-response on original microfilm.
NR^3 - Total non-response.

Field	Processing error	Other discrepancies(1)				
	Type unknown	Coding	CP	R	NR[1]	NR[3]
	per cent					
Provinces reporting on tape:						
Sex	.09	-	-	-	.34	-
Date of birth	.09	.09	-	.22	.17	-
Kind of birth	-	-	-	-	-	-
Residence - Mother	.09	.84	-	6.51	.26	.60
Duration of pregnancy	.09	-	12.03	-	.43	-
Total children born	18.51	2.23	-	-	-	-
Number children liveborn	2.49	.45	-	-	.09	-
Number children stillborn	16.34	6.90	-	-	-	-
Weight of child	.68	.22	7.35	.45	.17	1.71
Marital status - Mother	.09	-	-	-	.17	.09
Birthdate - Mother	.22	-	-	-	-	-
Age - Mother	.17	-	-	-	-	-
Total	**38.86**	**10.73**	**19.38**	**7.18**	**1.63**	**2.40**
Provinces reporting on micro-film/forms:						
Sex	-	-	-	-	-	-
Date of birth	-	-	-	.08	-	-
Kind of birth	.08	-	-	-	.25	-
Order of birth	-	-	-	-	.08	-
Residence - Mother	.08	1.80	.23	2.41	-	-
Total children born	8.88	-	-	-	-	-
Number children liveborn	.23	-	-	-	-	-
Number children stillborn	1.17	-	-	-	-	-
Weight of child	.47	-	-	-	-	-
Parents married to each other	-	-	-	-	-	-
Marital status - Mother	.47	-	-	-	-	-
Brithdate - Mother	-	-	-	-	-	-
Total	**11.38**	**1.80**	**.23**	**2.49**	**.33**	**-**
All provinces:						
Sex	.08	-	-	-	.30	-
Date of birth	.08	.08	-	.21	.15	-
Kind of birth	.01	-	-	-	.03	-
Order of birth	-	-	-	-	.01	-
Residence - Mother	.09	.94	.03	6.06	.23	.53
Duration of pregnancy	.08	-	10.70	-	.38	-
Total children born	17.45	1.98	-	-	-	-
Number children liveborn	2.25	.40	-	-	.08	-
Number of children stillborn	14.66	6.14	-	-	-	-
Weight of child	.66	.20	6.54	.40	.15	1.52
Parents married to each other	-	-	-	-	-	-
Marital status - Mother	.13	-	-	-	.15	.08
Birthdate - Mother	.20	-	-	-	-	-
Age - Mother	.15	-	-	-	-	-
Total	**35.84**	**9.74**	**17.27**	**6.67**	**1.48**	**2.13**

(1) CP - Coded by province from other information.
R - Reporting problem.
NR[1] - Non-response on original microfilm.
NR[3] - Total non-response.

TEXT TABLE III. Stillbirths - Weighted Results

Field	Processing error	Other discrepancies(1)					
	Type unknown	Coding	A	R	NR[1]	NR[3]	IMP
	per cent						
Provinces reporting on microfilm/forms:							
Sex	-	-	-	-	.47	-	-
Date of stillbirth	.40	-	-	2.46	-	-	-
Kind of birth	.40	1.20	-	-	-	-	-
Order of birth	-	-	-	-	.47	-	-
Residence of mother	-	1.26	.93	3.86	1.99	-	-
Duration pregnancy	1.26	-	-	.47	-	-	-
Total children born	2.80	-	.47	1.87	-	-	-
Number children liveborn	1.40	-	.47	.47	-	-	-
Number children stillborn	.93	-	.47	.93	-	-	-
Weight at birth	.47	.93	-	.47	-	3.73	-
Marital status - Mother	-	-	.47	-	-	-	79.15
Parents married to each other	.47	-	-	-	-	-	25.12
Birthplace - Father	.40	-	-	2.33	-	-	-
Birthdate - Mother	-	-	-	.47	-	-	25.21
Age - Mother	-	-	-	-	-	-	-
Cause of death	-	16.39	-	.80	-	-	-
Total	**8.53**	**19.78**	**2.81**	**14.13**	**2.93**	**3.73**	**129.48**

(1) A - Amendment.
R - Reporting problem.
NR[1] - Non-response on microfilm.
NR[3] - Total non-response.
IMP - Imputation.

TEXT TABLE IV. Deaths - Weighted Results

Field	Processing error	Other discrepancies(1)					
	Type unknown	Coding	A	CP	R	NR[1]	IMP	ILL
	per cent							
Provinces reporting on tape:								
Sex	.09	-	-	-	-	.17	-	-
Place of death	1.08	4.80	1.17	-	.90	1.80	-	.69
Residence - Deceased	-	9.65	.30	.22	1.94	.69	-	.43
Marital status	.60	-	.22	-	.47	.09	-	-
Birthdate	2.91	.22	-	.17	.17	.17	12.58	.47
Age	1.29	.22	.39	-	.43	1.54	-	.34
Date of death	.09	.09	-	-	.09	1.63	11.71	.82
Cause of death	-	7.01	-	-	-	-	-	.34
Autopsy	3.69	-	-	-	-	-	-	.09
Date of certificate	.22	-	-	-	-	-	-	.22
Total	**9.97**	**21.99**	**2.08**	**.39**	**4.00**	**6.09**	**24.29**	**3.40**
Provinces reporting on microfilm/forms:								
Sex	.25	-	-	-	-	-	-	.07
Place of death	-	3.23	.25	-	2.14	.13	-	.39
Residence - Deceased	-	3.16	1.50	-	.59	-	-	.13
Marital status	.20	-	-	-	.07	-	-	.13
Birthdate	.52	.25	1.00	-	.32	-	1.69	.32
Age	1.19	-	.50	-	.07	-	-	.13
Date of death	.58	.25	.32	-	-	-	-	.20
Cause of death	1.64	8.73	.25	-	-	-	-	.72
Autopsy	.25	-	-	-	-	-	-	.07
Date of certificate	.38	-	-	-	-	-	-	.26
Total	**5.01**	**15.62**	**3.82**	**-**	**3.19**	**.13**	**1.69**	**2.42**
All provinces:								
Sex	.10	-	-	-	-	.15	-	.01
Place of death	.97	4.64	1.08	-	1.03	1.62	-	.66
Residence - Deceased	-	9.00	.42	.19	1.81	.63	-	.40
Marital status	.56	-	.19	-	.43	.08	-	.01
Birthdate	2.67	.22	.10	.15	.19	.15	11.48	.43
Age	1.28	.19	.40	-	.39	1.39	-	.32
Date of death	.14	.10	.03	-	.08	1.46	10.56	.75
Cause of death	.17	7.18	.03	-	-	-	-	.38
Autopsy	3.35	-	-	-	-	-	-	.08
Date of certificate	.23	-	-	-	-	-	-	.22
Total	**9.47**	**21.33**	**2.25**	**.34**	**3.93**	**5.48**	**22.04**	**3.26**

(1) A - Amendment.
CP - Coded by province from other information.
R - Reporting problem.
NR[1] - Non-response on microfilm.
IMP - Imputation.
ILL - Illegible.

Number/
Numéro

2

Statistics Canada / Statistique Canada

Health Division / Division de la santé

Statistique Canada

Division de la santé

Évaluation de la qualité des statistiques de l'état civil

(Étude - Pilote)

par

D.N. Nagnur
S.G. Currie
B. Heath

Publication autorisée par
le ministre des
Approvisionnements et
Services Canada

Reproduction ou citation
autorisées sous réserve
d'indication de la source:
Statistique Canada

© Ministre des Approvisionnements
et Services Canada 1981

Août 1981
4-2302-535

Ottawa

This publication is available
in English upon request.

ÉVALUATION DE LA QUALITÉ DES STATISTIQUES DE L'ÉTAT CIVIL - ÉTUDE-PILOTE

La qualité des statistiques de l'état civil publiées par Statistique Canada est censée avoir toujours été très élevée, mais il n'a jamais été tenté de procéder à une évaluation quantitative de la qualité des données avant la présente étude. Lorsqu'il a été entrepris, le projet portait le nom "d'Évaluation du contenu et de la qualité des statistiques de l'état civil intégrées". L'étude portait sur les naissances, les mortinaissances, les mariages et les décès survenus au cours de 1976.

Comme chacun des événements démographiques doit, aux termes de la loi, être enregistré, on a supposé, à juste titre, que les événements enregistrés étaient exhaustifs.

Les provinces et les territoires recueillent les statistiques de l'état civil au moyen de formules d'enregistrement conçues pour répondre aux impératifs administratifs et en conformité de la loi. En vertu d'ententes fédérales-provinciales qui remontent à 1919, Statistique Canada utilisent ces documents pour produire des données nationales sur la mortalité, la natalité, la nuptialité et la longévité.

Les ministères, les organismes privés et le grand public se servent de plus en plus des statistiques de l'état civil pour procéder à la planification et prendre des décisions en matière de politique - notamment en ce qui concerne les domaines de la santé et des programmes sociaux. Par conséquent, l'analyse et la mesure des tendances démographiques et épidémiologiques que traduisent les données chronologiques sur les événements démographiques revêtent une importance et un caractère immédiat accrus.

Pour évaluer la qualité des statistiques de l'état civil, on peut procéder de plusieurs façons. On s'est cependant servi de l'ordre suivi pour produire les statistiques à partir de la collecte initiale des données. Voici quelques-unes des étapes importantes à étudier:

a) le délai d'enregistrement et la qualité des données déclarées par chacun des intéressés;

b) la vérification initiale et l'acheminement des formules d'enregistrement du réseau des centres régionaux aux bureaux provinciaux et territoriaux;

c) le retard imputable au processus de transmission des données et les opérations de traduction des données sous forme ordinolingue;

d) les modes de transformation des données sur le support de la Section de la statistique de l'état civil et des registres des maladies de Statistique Canada et l'effet que cette transformation a sur le délai, la qualité, etc.;

e) la fréquence et le délai de modification des documents originaux au palier provincial;

f) après la réception des données au bureau de la statistique de l'état civil de Statistique Canada, les données passent par de multiples étapes comme la vérification, l'imputation, etc. et permettent de produire une foule de totalisations. On a recours à des programmes informatiques pour certaines des opérations et à des manipulations manuelles pour d'autres.

Afin d'obtenir un profil complet de la qualité des statistiques de l'état civil, il est essentiel d'analyser chacune des étapes suivies pour la traduction des "données" en "statistiques". Cependant, plusieurs contraintes nous obligent à nous restreindre à un petit sous-ensemble de l'univers. La première étape consiste dans la réception, à la Section de la statistique de l'état civil et des registres des maladies de Statistique Canada, des données provenant des bureaux provinciaux et territoriaux de la statistique de l'état civil. La dernière étape consiste dans l'enregistrement des données ayant subi une vérification finale à l'aide d'un système informatisé utilisé pour la production de totalisations et de rapports.

La méthode d'échantillonnage établie et adoptée comportait une stratification fondée surtout sur le mode d'obtention des données des provinces qui devaient être dépouillées et totalisées. L'échantillon comprenait 4,400 documents distribués entre toutes les provinces, soit 1,300 pour chacun des événements suivants: naissance, mariage et décès, et 500, pour les mortinaissances. Cet échantillon représentait 0.2% des naissances annuelles, 0.7% des mariages, 0.8% des décès et environ 19% des mortinaissances. L'échantillon global de tous les événements combinés correspondait à environ 0.6% de tous les événements survenus en 1976. Seuls les domaines clés représentant ce que l'on a appelé les caractéristiques principales dans chaque série ont fait l'objet d'un examen détaillé.

Les caractéristiques principales étudiées en ce qui concerne les différents événements étaient les suivantes:

Naissances

Sexe de l'enfant
Date de naissance
Genre d'accouchement
Rang de naissance
Lieu de résidence de la mère
Durée de la grossesse
Nombre d'enfants nés
Enfants nés vivants
Enfants mort-nés
Poids à la naissance de l'enfant
Parents mariés l'un à l'autre
État matrimonial de la mère
Date de naissance de la mère
Âge de la mère

Mortinaissances

Sexe de l'enfant
Date de mortinaissance
Genre d'accouchement
Rang de naissance
Lieu de résidence de la mère
Durée de la grossesse
Nombre d'enfants nés
Nombre d'enfants nés vivants
Nombre d'enfants mort-nés
Poids à la naissance
État matrimonial de la mère
Parents mariés l'un à l'autre
Lieu de naissance de la mère
Lieu de naissance du père
Date de naissance de la mère
Âge de la mère
Cause de la mortinaissance

Mariages

Date du mariage
État matrimonial de l'époux
Date de naissance de l'époux
Âge de l'époux
État matrimonial de l'épouse
Date de naissance de l'épouse
Âge de l'épouse

Décès

Sexe
Lieu de décès
Lieu de résidence de la personne décédée
État matrimonial
Date de naissance
Âge
Date du décès
Cause du décès
Autopsie
Date de certification

Il avait d'abord été proposé d'avoir recours à un appariement triple. Les données initiales reçues des provinces devaient être appariées aux enregistrements recodés à partir des exemplaires microfilmés des documents originaux et à ceux de la bande finale vérifiée. En raison de difficultés inévitables, il a fallu s'en tenir à un appariement double - les enregistrements de la bande finale vérifiée ont été appariés aux enregistrements recodés provenant des documents microfilmés originaux. Il est prévu de procéder à un appariement triple dans les études qui auront lieu au cours des prochaines années.

Résultats

Les résultats indiquent que le niveau de précision et de fiabilité des diverses phases de dépouillement interne à la Section de la statistique de l'état civil et des registres des maladies de Statistique Canada est acceptable. Le taux d'erreur mesuré en fonction des "mauvais appariements" attribués à la majorité des caractéristiques principales indique que le degré de précision et d'uniformité est raisonnable.

Le rapport établi par S.G. Currie de la Division de la méthodologie donne la ventilation détaillée des diverses caractéristiques. Ce rapport figure à l'appendice 1 ci-joint. On trouvera ci-après un bref résumé des résultats de chaque série.

Mariages

Le taux d'erreur global pour les sept caractéristiques principales des mariages s'établissait à 1.33 pour 100 documents, et aucune des caractéristiques n'affichait un taux d'erreur supérieur à 0.5 pour 100 documents. Parmi les caractéristiques dont le taux d'erreur était relativement élevé figuraient la date de naissance, l'état matrimonial et l'âge de l'époux; le taux d'erreur des deux dernières était moins élevé, environ 0.25 pour 100 documents.

En outre, il n'y avait aucune différence importante entre les provinces qui ont présenté leurs données sur bande et celles qui ont fourni des microfilms ou des formules. On a aussi constaté qu'une partie appréciable des incompatibilités était attribuable au manque de renseignements et à l'imputation ultérieure.

Naissances

Le taux d'erreur global de toutes les caractéristiques principales des naissances, sauf celles ayant trait au nombre d'enfants nés, s'établissait à 2.7 pour 100 documents. Le taux d'erreur le plus élevé, soit pour le "lieu de résidence de la mère", était d'environ 1%. Venait ensuite le "poids à la naissance de l'enfant", dont le taux se chiffrait à 0.9 pour 100 documents. Les taux de non-réponse de ces deux caractéristiques s'établissaient également à 0.5% et 1.5% respectivement. Il faut étudier davantage ces

caractéristiques. Comme la fréquence des non-réponses devrait diminuer, il est probable que le taux d'erreur décroîtra encore davantage.

Parmi les autres caractéristiques qui ont connu des incompatibilités importantes figurent le "nombre total d'enfants", "le nombre d'enfants mort-nés" et, dans une moindre mesure, le "nombre d'enfants nés vivants". Les incompatibilités ne se traduisent toutefois pas dans la même mesure par des erreurs dans les totalisations. Certaines incompatibilités semblent découler du "libellé" des questions sur le nombre d'enfants. Les provinces ont recours à un libellé différent. Une partie du problème peut aussi être attribuée aux données manquantes et à l'imputation ultérieure. Il faudra examiner davantage ces aspects au cours des études ultérieures.

Le taux d'erreur des provinces qui se sont servies de microfilms ou de formules n'est que légèrement inférieur à celui des provinces qui ont utilisé des bandes. Pour isoler davantage les erreurs, on surveillera avec soin le codage et d'autres aspects, notamment pour les caractéristiques comme le "lieu de résidence de la mère", le "poids à la naissance de l'enfant" et la "durée de la grossesse" au cours des prochaines études.

Décès

Il faut interpréter avec prudence les résultats des "décès", car, même si les taux d'erreur tant individuels que globaux semblent appréciables, les effets que ces derniers ont sur les principales totalisations ne correspondent pas à leur importance.

Par exemple, le taux d'erreur de plus de trois pour 100 documents qu'affichent les autopsies n'a aucun effet sur les totalisations relatives à la mortalité produites normalement. De même, le "lieu de décès", qui accuse un taux d'erreur supérieur à cinq pour 100 documents, n'a pas d'effet sur les totalisations, car toutes les totalisations relatives aux "décès" sont fondées sur le "lieu normal de résidence". Le taux élevé d'erreur du "lieu de résidence du décédé" aurait un effet, mais l'analyse détaillée révèle qu'il se produit surtout au "niveau de la localité" et qu'aux niveaux du comté, de la division de recensement ou à un niveau plus élevé, le taux d'erreur diminue de façon sensible. Comme la plupart des totalisations sur la mortalité s'appliquent au niveau national ou provincial, le taux d'erreur du "lieu de résidence" révélé par l'étude est moins grave qu'il ne le semble en réalité. Le taux d'erreur de la "cause de décès" suscite plus de préoccupations; c'est pourquoi il semble nécessaire de surveiller attentivement les sources de ce taux d'erreur afin de pouvoir les réduire au minimum et de les corriger. Cela est d'autant plus important que cette classification sert de base à de nombreuses études de postobservation épidémiologiques et médicales. Compte tenu de ce fait, l'étude suivante portera sur la cause de décès pour 1979. Les résultats de cette étude devraient jeter plus de lumière sur les erreurs qui s'appliquent à la cause de décès.

Les taux d'erreur globaux des décès peuvent être considérés comme légèrement supérieurs à ceux des "mariages" et des "naissances"; mais cela peut s'expliquer par le fait que les données sur la personne décédée doivent nécessairement être fournies par les parents et d'autres personnes et que les modifications aux données initiales sont évidemment plus fréquentes. L'imputation de certaines caractéristiques comme la "date de naissance" augmente le taux global d'erreur même si cela n'a pas d'effet important lorsque l'âge est déclaré ailleurs. Aux fins des totalisations, plusieurs aspects des données sur le décès seront examinés au cours des prochaines études.

Mortinaissances

L'exhaustivité du champ d'observation des mortinaissances revêt une grande importance, mais leurs caractéristiques sont peut-être moins importantes par rapport aux naissances, aux mariages et aux décès. La mortalité périnatale est l'indicateur le plus important pour lequel le nombre de mortinaissances joue un rôle primordial. Cependant d'autres caractéristiques des mortinaissances revêtent une grande importance pour l'étude de la tendance des décès des foetus de 28 semaines et plus, leurs causes et leurs conséquences.

La plupart des erreurs observées concernant les "mortinaissances" avaient trait au "nombre d'enfants" et à la "cause du décès". Le nombre d'enfants nés influe sur les totalisations relatives aux naissances selon le rang de naissance de tous les enfants, et lorsqu'on examine la répartition des naissances selon le rang de naissance, il faut préférer la répartition fondée sur le rang de naissance des enfants vivants plutôt que sur le rang de naissance de tous les enfants. Il faut également s'efforcer d'améliorer les statistiques sur le nombre d'enfants, nés vivants ou mort-nés.

La cause de mortinaissance est une autre caractéristique dont le niveau d'erreur est appréciable. L'examen détaillé révèle que le taux d'erreur est moins important que ne l'indiquent les résultats, puisque plus de la moitié des cas ayant contribué à l'erreur, correspondaient aux deux premiers chiffres des rubriques à quatre chiffres. Le lieu de résidence de la mère, comme c'était le cas pour les autres événements, différait de façon plus sensible au niveau de la localité et au niveau global de comté ou de la division de recensement et à un niveau supérieur; cette erreur n'a pas constitué un facteur très important.

Conclusion générale

La conclusion générale à tirer de l'étude est que l'incidence des erreurs qui découlent du dépouillement interne n'est pas, de façon générale, grave bien qu'il y ait des domaines et des caractéristiques où il y a lieu et où il est souhaitable d'apporter des améliorations qui permettraient sans aucun doute d'obtenir et de diffuser des renseignements de meilleure qualité. Il existe cependant des différences marquées d'une région à l'autre dans la qualité des données concernant certaines caractéristiques et certains événements; l'un des facteurs de cet état de fait est le manque de renseignements ou les non-réponses. On pourrait corriger cette situation et peut-être l'améliorer en surveillant de près les données au stade de la collecte ou du dépouillement et en établissant un système de rappel. L'étude fait également ressortir le fait qu'il faudrait procéder à un examen détaillé à l'aide de l'appariement triple proposé au début et en accroissant la taille de l'échantillon pour fournir des mesures fiables au niveau provincial. Afin d'obtenir un profil complet de la qualité des statistiques de l'état civil, il faut entreprendre l'analyse de chacune des étapes suivies pour la traduction des "données" en "statistiques" comme il a été proposé plus haut. L'étude sur l'évaluation de la qualité doit logiquement mener à l'établissement d'un mécanisme permettant de contrôler la qualité et de communiquer avec les organismes répondants des provinces pour obtenir des renseignements sur les données de l'année en cours. Une fois terminées les études-repères visant à établir le modèle des mesures de fiabilité, à l'échelle nationale et provinciale, pour différents événements démographiques répartis selon les caractéristiques principales, les prochaines études pourront être axées sur les aspects spécialisés du système de statistiques de l'état civil.

À titre de complément de la présente étude sur l'évaluation de la qualité, des formules statistiques servant à mesurer la variance et les erreurs-types des taux de mortalité et de fécondité les plus souvent employés ont fait l'objet d'un examen. Une note technique à ce sujet peut être obtenu sur demande. Pour améliorer le contenu et la qualité des rapports annuels sur les naissances et les décès, nous ferons figurer certaines des mesures de la variance et des erreurs-types, probablement pour l'année 1979.

APPENDICE I

ÉVALUATION DU DÉPOUILLEMENT DES STATISTIQUES DE L'ÉTAT CIVIL

MARIAGES, NAISSANCES, MORTINAISSANCES ET DÉCÈS

La qualité des statistiques de l'état civil a toujours été considérée comme bonne même si l'on n'a effectué aucune évaluation directe avant la présente étude. L'évaluation de la qualité des données au moyen d'études de ce genre est conforme à la politique qu'a adoptée récemment le bureau en matière de qualité des données. Dans la mesure où les résultats de la présente évaluation et des évaluations futures fourniront des mesures significatives de la qualité des données, ils paraîtront dans les publications de la statistique de l'état civil. Les résultats peuvent également servir de base à l'amélioration méthodologique de la collecte et du dépouillement des données.

La présente étude a pour but de mesurer les taux d'erreur produits par les opérations de codage, de perforation et de vérification au cours du traitement des statistiques de l'état civil. Comme les provinces effectuent maintenant la plupart des travaux de codage des données et de perforation, la nécessité d'établir un système de surveillance pour faire en sorte que les normes de qualité soient respectées est devenu un sujet de préoccupation pour Statistique Canada. La présente étude ne doit pas être considérée comme une tentative ponctuelle visant à mesurer la qualité des données, mais plutôt comme le point de départ de l'établissement d'un système permanent de surveillance de la qualité des statistiques de l'état civil.

Méthodologie

L'évaluation des taux d'erreur de dépouillement est fondée sur un échantillon de documents de chaque série qui ont fait l'objet d'un deuxième dépouillement indépendant et qui ont été comparés aux documents originaux de la base de données épurée. Seules les principales caractéristiques ont été examinées.

On s'attendait à ce que les taux d'erreur varient d'une province à l'autre, car les provinces procèdent à divers travaux de dépouillement à l'aide de formules et de méthodes légèrement différentes. Les échantillons ont été choisis de la façon suivante:

	Naissances	Mariages	Mortinaissances	Décès
Provinces présentant leurs déclarations sur des bandes:				
Québec	326	329	-	334
Autres	324	322	287	329
Provinces présentant leurs déclarations sur des microfilms ou des formules:				
Terre-Neuve	319	305	74	300
Autres	313	318	151	322
Total	1,282	1,274	512	1,285

Les échantillons du Québec et de Terre-Neuve ont été choisis séparément, car on était d'avis que ces deux provinces pourraient présenter des problèmes particuliers. Les échantillons des deux groupes de provinces ont été alloués à chaque province proportionnellement.

On a choisi les échantillons à l'aide des numéros d'enregistrement plutôt qu'à partir des microfilms ou des bandes d'ordinateur. L'étendue des numéros d'enregistrement a été

déterminée pour chaque province, et ces renseignements ainsi que la taille de l'échantillon désirée ont été introduits dans un programme de sélection de l'échantillon qui a produit une liste de numéros d'enregistrement de l'échantillon. Les enregistrements correspondant aux numéros d'enregistrement choisis ont été repérés sur les photocopies et les microfilms produits.

Comme le codage original des enregistrements figurait sur les microfilms, il a fallu effacer les codes sur les photocopies afin de rendre le recodage indépendant de l'original. Il est extrêmement important que le recodage soit effectué de façon indépendante pour obtenir des taux d'erreur précis, car on sait que les recodeurs ont tendance à accepter le code original plus souvent s'ils le connaissent. Les recodeurs ont enregistré les codes sur des feuilles spécialement conçues à cet effet qu'ils ont ensuite perforées. La totalité des données introduites par perforation ont été vérifiées pour faire en sorte que les erreurs de perforation soient réduites au minimum.

Selon la méthodologie originale, il fallait que les enregistrements recodés soient appariés aux bandes de données brutes ainsi qu'au fichier vérifié épuré. Malheureusement, il n'était pas facile d'avoir accès aux bandes de données brutes à cette fin et il a fallu les supprimer de l'appariement. Cela a donc rendu l'identification des genres d'erreurs beaucoup plus difficile.

On a apparié les enregistrements recodés et les enregistrements correspondants du fichier vérifié épuré et on a produit une liste de toutes les différences. Une personne a examiné les différences et attribué les erreurs. Malheureusement, cette dernière n'avait pas accès aux renseignements corrigés, car il aurait fallu trop de temps pour les extraire. C'est pourquoi, les jugements portés sont fondés sur les enregistrements des microfilms et la liste des modifications.

On a pondéré les résultats de l'étude afin de combiner les estimations à partir des strates de l'échantillon. Il a fallu procéder ainsi car les proportions d'échantillonnage variaient considérablement. Le poids de la i^e strate était le suivant:

$$Wi = \frac{Ni}{\sum_{i} Ni}$$

où N_i est le nombre total d'enregistrements de la i^e strate. L'estimation combinée du taux d'erreur provient de

$$ec = \frac{\sum Wi.ei}{\sum_{i} Wi}$$

où e_i est le taux d'erreur de l'échantillon à partir de la i^e strate. Il convient de noter que $\sum_{i} Wi=1$ pour les estimations nationales, mais serait inférieur à 1 lorsqu'on estime le taux d'erreur pour les provinces qui présentent les déclarations sur des bandes. Les valeurs de N_i utilisées pour le calcul des poids de la présente étude sont préliminaires et ne correspondent pas exactement aux poids calculés à partir des totaux publiés.

Résultats des mariages

Les résultats ont été très satisfaisants pour chacune des caractéristiques qui ont été vérifiées. Cependant, certaines lacunes des données ne se traduisent pas dans les résultats de l'échantillon. La Colombie-Britannique et l'Ontario ne déclarent pas les dates de naissance. Comme les taux d'erreur pour la date de naissance sont plus élevés que ceux des autres caractéristiques principales, les résultats de ces provinces semblent meilleurs qu'ils le sont en réalité. De même, la Colombie-Britannique ne déclare pas le jour du mariage sur ses bandes (seulement le mois et l'année) même si le jour figure sur le microfilm.

Il n'a pas été possible d'identifier les erreurs de perforation au cours du dépouillement original, mais selon les résultats du retraitement, la perforation n'est probablement pas un problème sérieux. La seule erreur de perforation identifiée était attribuable à un seul enregistrement de mariage où les données avaient été déplacées d'une case, ce qui avait causé des erreurs dans trois zones.

La méthode d'imputation de l'âge à partir de la date de naissance est un autre problème que ne reflètent pas les résultats de l'échantillon. Le programme d'imputation informatisé calcule l'âge à partir de l'année de naissance et de l'année du mariage plutôt qu'au moyen des dates complètes. Il en est résulté de nombreuses incompatibilités dans l'échantillon qui n'ont pas été comptées comme des erreurs. Par exemple, au Manitoba,des incompatibilités se sont produites pour environ la moitié des échantillons d'enregistrements.

Des non-réponses se sont produites lorsque la date de naissance et l'âge étaient laissés en blanc. Il convient de noter que si la date de naissance est déclarée, mais non l'âge, l'âge est imputé à partir de la date de naissance dans le programme de contrôle. Pour les microfilms, c'est surtout l'état matrimonial qui a fait l'objet de non-réponses. Dans ces cas, l'état matrimonial a été considéré comme étant le célibat s'il n'y avait pas d'indication d'un mariage précédent.

La comparaison des taux d'erreur de dépouillement des données des provinces qui présentent leurs déclarations sur bandes et de

celles qui utilisent des microfilms ou des formules ne révèle pas des différences importantes.

Naissances

La principale caractéristique en ce qui concerne les naissances est le taux élevé d'erreur pour les trois zones relatives au "nombre d'enfants nés de cette mère". Si l'on inclut ces zones, le taux d'erreur pour toutes les caractéristiques s'établissait à 45.55 erreurs pour 100 documents. Mais si on les exclut, le taux d'erreur tombe à 2.70 erreurs pour 100 documents, ce qui est très raisonnable.

La plupart des erreurs concernant les trois zones relatives au "nombre d'enfants nés de cette mère" n'étaient pas réellement des erreurs de codage au sens strict du mot, mais des différences entre un code et une zone laissés en blanc. Il s'agit essentiellement de savoir comment traiter ces zones lorsqu'elles sont laissées en blanc au départ. Il est possible de les remplir par une correction ou par une imputation. Il est probable qu'on a imputé une valeur plutôt que de recueillir plus de renseignements par rappel. Les imputations les plus courantes sont les suivantes:

(1) si le nombre de mort-nés et le nombre total d'enfants nés ne sont pas indiqués, on impute le chiffre 0 pour les mort-nés et le nombre total d'enfants nés équivaut au nombre d'enfants nés vivants;

(2) si le nombre de nés vivants n'est pas indiqué, on impute le chiffre 1 à la rubrique "nés vivants".

Il est tout à fait possible que ces imputations causent certains biais, mais on ne peut tirer aucune conclusion certaine d'après la présente étude.

Pour les autres zones, le taux d'incompatibilités était élevé, car on a attribué le code à l'aide de renseignements qui ne figuraient pas sur le microfilm original. Ces incompatibilités se sont produites surtout dans le cas de deux caractéristiques, soit la "durée de la grossesse" et le "poids de l'enfant". Pour l'Ontario et la Colombie-Britannique, on a changé la "durée de la grossesse" d'une façon relativement constante - on a ajusté les valeurs faibles à la hausse et les valeurs élevées ont été fréquemment changées. Il s'agit de déterminer si des renseignements supplémentaires sont obtenus dans ces cas ou si des valeurs aient été modifiées d'une façon déterminée au préalable.

La caractéristique "poids de l'enfant" a causé de nombreuses incompatibilités pour l'Ontario, parce que cette province arrondit toujours les onces à la fraction supérieure tandis que les recodeurs les arrondissent à la fraction supérieure ou inférieure de la façon habituelle. Il en découle donc un léger biais à la hausse concernant le "poids de l'enfant" qui n'est probablement pas important.

Dans certains cas, on n'a pas indiqué la "durée de la grossesse" et "poids de l'enfant", car 0.38% et 0.15% des cas ne figuraient pas sur le microfilm, mais étaient codés sur le fichier vérifié tandis que pour le "poids de l'enfant", il n'y avait pas de réponse dans 1.52% des cas.

À l'exclusion du "nombre d'enfants nés", la caractéristique qui a donnée lieu au plus grand nombre d'erreurs était le "lieu de résidence de la mère", dont le taux d'erreur s'établissait à 1.03%. En outre, le taux d'incompatibilités s'élevait à 6.85%, dont la plupart étaient attribuables à de mauvaises déclarations, y compris les réponses illisibles et illogiques ainsi que les non-réponses. L'utilisation d'un manuel de codage qui n'était pas à jour a également causé quelques problèmes.

Les provinces qui ont présenté leurs déclarations sur des microfilms ou des formules affichaient des taux d'erreur moins élevés que celles qui ont utilisé des bandes.

Résultats des mortinaissances

En ce qui concerne les mortinaissances, le taux d'erreur de traitement pour 100 documents s'établissait à 13.10 pour les trois zones du "nombre d'enfants nés", à 13.75 pour la "cause du décès", à 4.29 pour le "lieu de résidence de la mère" et à 3.16 pour les autres caractéristiques.

Les taux d'erreur du "nombre d'enfants nés" sont moins élevés que ceux des naissances, mais cela reflète peut-être davantage la variabilité du processus d'attribution d'une différence réelle. Il ne fait aucun doute que ces éléments sont ambigus, surtout si l'on tient compte du fait qu'à la suite de changements du libellé de certaines questions, on ne savait plus si l'on devait déclarer l'événement ou non. La déclaration des naissances multiples a prêté à confusion et l'on ne pourra régler ce problème qu'en donnant des directives plus claires aux personnes qui remplissent les formules.

Le taux d'erreur de 4.29% pour le "lieu de résidence de la mère" était supérieur à celui des naissances, mais il était acceptable. Il convient toutefois de noter que les erreurs de réponse représentent 7.34% des incompatibilités. Cela reflète probablement une déclaration moins que précise de la municipalité et, dans une certaine mesure, des problèmes de manuel de codage.

Les non-réponses s'appliquaient à deux caractéristiques. Le taux de non-réponse du "sexe" s'établissait à 1.24% pour les

microfilms; il semble que les données figurant dans le fichier vérifié aient été imputées manuellement au cours de la correction. Le "poids à la naissance" était laissé en blanc sur le microfilm et le fichier vérifié dans 2.65% des cas. L'absence du poids à la naissance dans le cas d'une mortinaissance est importante, car le poids ainsi que la durée de la grossesse servent à déterminer la nature de l'événement, c.-à-d., à faire la distinction entre une fausse couche et un avortement.

Il convient de noter que, même si le taux d'erreur pour le "poids à la naissance" était faible, le taux de correction de 6.72% était élevé. La presque totalité des modifications s'appliquaient aux données de l'Ontario.

Les deux caractéristiques "parents mariés l'un à l'autre" et "état matrimonial de la mère" ont donné lieu à des taux d'imputation élevés de 13.11% et 41.69% respectivement. Le système impute "oui" pour les "parents mariés l'un à l'autre" si l'"état matrimonial de la mère" est "mariée" et vice versa. Cette façon de procéder laisse à désirer, car la mère peut, dans certains cas, être mariée à quelqu'un d'autre que le père.

On a imputé la "date de naissance de la mère" un nombre considérable de fois, soit dans 56.09% des cas. Dans ces cas, le système informatique a créé une "année de naissance" plutôt qu'une "date de naissance" à l'aide de l'âge de la mère. Il faut examiner l'utilité de cette façon de procéder et sa précision. Le taux d'erreur de codage pour la cause de décès était assez élevé 13.75%, beaucoup plus élevé que le taux d'erreur de la série sur les décès. On peut peut-être s'y attendre, car la cause de décès pour les mortinaissances peut être plus mal définie que dans le cas des autres décès.

Une analyse indique que, pour 31.9% des erreurs, il y avait correspondance, sauf concernant le 4^{e} et dernier chiffre du code, et dans 24.6% des cas, les deux premiers chiffres correspondaient. On peut en déduire que le taux d'erreur est fonction du niveau de détail du codage.

Résultats des décès

En général, les résultats des décès ne sont pas aussi bons que ceux des autres séries. Il est évident que le codage géographique (lieu du décès et lieu de résidence de la personne décédée) ont causé le principal problème. Le taux d'erreur de 9.00% pour le "lieu de résidence de la personne décédée" est supérieur aux taux d'erreur du "lieu de résidence de la mère" dans les fichiers des naissances et des mortinaissances. Le taux d'erreur de 5.61% pour le "lieu de décès" est également élevé pour un élément d'information aussi important. Il semble que le piètre codage géographique est attribuable à de mauvais manuels de codage et à la déclaration imprécise des lieux géographiques.

Il vaut la peine de noter les caractéristiques "date de naissance" et "autopsie" dont les taux d'erreur s'établissaient à 2.89% et 3.25% respectivement. Ce sont des taux relativement élevés pour des zones simples. On ne sait pas quelle proportion des taux d'erreur est attribuable au codage et à la perforation, mais il est clair que les taux d'erreur des caractéristiques semblables des autres genres de documents étaient inférieurs.

Les imputations ont touché deux caractéristiques. Le taux d'imputation de 11.48% pour la "date de naissance" provient du calcul de la date de naissance à partir de l'âge lorsque la date de naissance n'était pas déclarée. Pour la "date du décès", le taux d'imputation s'établissait à 10.56%. Ce phénomène est dû à la non-déclaration du "jour du décès" dans certaines provinces auquel le système informatique impute "01".

Le taux d'erreur de codage de 7.18% pour la "cause du décès" semble assez raisonnable, bien qu'il soit possible de faire mieux si l'on exerce un contrôle strict de cette opération de codage. Environ les deux tiers des erreurs concernaient le 1er et le 2^{e} chiffre du code à quatre chiffres. Cela indique que les erreurs étaient plus graves que celles qu'on a rencontrées dans le codage des mortinaissances.

Conclusions et recommandations

Selon la présente étude, il semble que les taux d'erreur de dépouillement varient de très faibles pour les mariages à modérément élevés pour certaines caractéristiques des séries sur les mortinaissances et les décès; bien que les taux d'erreur ne soient probablement pas assez élevés pour susciter des préoccupations, ils sont assez élevés pour justifier d'autres examens. D'autre part, il a été prouvé à l'évidence que les non-réponses et les façons de les résoudre constituent un problème plus sérieux. Il semble que les données fassent l'objet de diverses modifications, corrections et imputations et qu'on ne soit pas certain que certaines de ces opérations n'introduisent pas des biais dans les données.

La présente étude est insuffisante dans la mesure où elle n'a pas pu identifier avec précision les sources d'erreurs. Si l'on avait disposé des données introduites par perforation à l'origine et des renseignements sur les corrections apportées, il aurait été possible de mieux évaluer la nature des erreurs de dépouillement constatées. Vu qu'on ne disposait pas des renseignements nécessaires pour prendre des décisions, la précision et l'uniformité du processus d'attribution laissaient peut-être à désirer.

Les recommandations suivantes sont formulées:

(1) examiner les questions sur le "nombre d'enfants nés de la mère" pour obtenir des réponses claires et cohérentes, surtout en ce qui concerne les accouchements multiples. Si c'est impossible, les statistiques sur le "rang de naissance" en souffriront énormément;

(2) réviser les méthodes de codage géographique et prendre des mesures pour que les codeurs des provinces aient des manuels de codage mis à jour le plus tôt possible;

(3) surveiller continuellement la qualité du codage de la "cause du décès" à l'aide de petits échantillons susceptibles d'être recodés par des codeurs experts;

(4) examiner la nature des modifications apportées aux données comme celles sur la "durée de la grossesse";

(5) examiner tous les programmes d'ajustement et d'imputation utilisés pour le dépouillement pour faire en sorte qu'ils ne produisent pas de biais graves;

(6) identifier et contrôler les causes des non-réponses. Lorsque certaines provinces ne recueillent pas certaines données clées, il faut entreprendre des négociations avec elles pour qu'elles les fassent figurer dans les formules au cours des prochaines années;

(7) entreprendre dans l'avenir une étude semblable à la présente étude à l'aide d'une méthodologie qui permettra de mieux identifier la source des erreurs.

TABLEAU EXPLICATIF I. Mariages - Résultats pondérés

Caractéristiques	Erreur de dépouillement	Autres incompatibilités(1)				
		CP	D	NR^1	NR^3	Total
	pourcentage					
Provinces utilisant des bandes:						
Date du mariage	.09	.09	.17	-	-	.26
État matrimonial - Époux	.26	-	-	.79	-	.79
Date de naissance - Époux	.44	-	.44	.09	-	.53
Âge - Époux	.26	-	.09	.09	.17	.35
État matrimonial - Épouse	.09	-	-	.35	-	.35
Date de naissance - Épouse	.09	-	.35	.09	-	.44
Âge - Épouse	.09	-	.22	.09	.17	.48
Total	**1.32**	**.09**	**1.27**	**1.50**	**.34**	**3.20**
Provinces utilisant des microfilms/formules:						
Date du mariage	.24	-	-	-	-	-
État matrimonial - Époux	.07	-	-	-	-	-
Date de naissance - Époux	.73	-	-	-	-	-
Âge - Époux	.24	.24	-	-	-	.24
État matrimonial - Épouse	.24	-	-	-	-	-
Date de naissance - Épouse	-	-	-	-	-	-
Âge - Épouse	-	-	-	-	-	-
Total	**1.32**	**.24**	**-**	**-**	**-**	**.24**
Toutes les provinces:						
Date du mariage	.10	.08	.16	-	-	.24
État matrimonial - Époux	.24	-	-	.71	-	.71
Date de naissance - Époux	.46	-	.40	.08	-	.48
Âge - Époux	.26	.02	.08	.08	.16	.34
État matrimonial - Épouse	.10	-	-	.32	-	.32
Date de naissance - Épouse	.08	-	.32	.08	-	.40
Âge - Épouse	.08	-	.20	.08	.16	.44
Total	**1.32**	**.10**	**1.16**	**1.35**	**.32**	**2.93**

(1) CP - Codé par la province à partir d'autres renseignements.
D - Problème de déclaration.
NR^1 - Non-réponse sur le microfilm original.
NR^3 - Total des non-réponses.

TABLEAU EXPLICATIF II. Naissances - Résultats pondérés

Caractéristiques	Erreur de dépouillement		Autres incompatibilités(1)			
	Genre inconnu	Codage	CP	D	NR[1]	NR[3]
	pourcentage					
Provinces utilisant des bandes:						
Sexe	.09	-	-	-	.34	-
Date de naissance	.09	.09	-	.22	.17	-
Genre - Accouchement	-	-	-	-	-	-
Résidence - Mère	.09	.84	-	6.51	.26	.60
Durée de la grossesse	.09	-	12.03	-	.43	-
Total - Enfants - Nés	18.51	2.23	-	-	-	-
Nombre d'enfants - Nés vivants	2.49	.45	-	-	.09	-
Nombre d'enfants - Mort-nés	16.34	6.90	-	-	-	-
Poids de l'enfant	.68	.22	7.35	.45	.17	1.71
État matrimonial - Mère	.09	-	-	-	.17	.09
Date de naissance - Mère	.22	-	-	-	-	-
Âge - Mère	.17	-	-	-	-	-
Total	**38.86**	**10.73**	**19.38**	**7.18**	**1.63**	**2.40**
Provinces utilisant des microfilms/formules:						
Sexe	-	-	-	-	-	-
Date de naissance	-	-	-	.08	-	-
Genre - Accouchement	.08	-	-	-	.25	-
Rang de naissance	-	-	-	-	.08	-
Résidence - Mère	.08	1.80	.23	2.41	-	-
Total - Enfants - Nés	8.88	-	-	-	-	-
Nombre d'enfants - Nés vivants	.23	-	-	-	-	-
Nombre d'enfants - Mort-nés	1.17	-	-	-	-	-
Poids de l'enfant	.47	-	-	-	-	-
Parents mariés l'un à l'autre	-	-	-	-	-	-
État matrimonial - Mère	.47	-	-	-	-	-
Date de naissance - Mère	-	-	-	-	-	-
Total	**11.38**	**1.80**	**.23**	**2.49**	**.33**	**-**
Toutes les provinces:						
Sexe	.08	-	-	-	.30	-
Date de naissance	.08	.08	-	.21	.15	-
Genre - Accouchement	.01	-	-	-	.03	-
Rang de naissance	-	-	-	-	.01	-
Résidence - Mère	.09	.94	.03	6.06	.23	.53
Durée de la grossesse	.08	-	10.70	-	.38	-
Total - Enfants - Nés	17.45	1.98	-	-	-	-
Nombre d'enfants - Nés vivants	2.25	.40	-	-	.08	-
Nombre d'enfants - Mort-nés	14.66	6.14	-	-	-	-
Poids de l'enfant	.66	.20	6.54	.40	.15	1.52
Parents mariés l'un à l'autre	-	-	-	-	-	-
État matrimonial - Mère	.13	-	-	-	.15	.08
Date de naissance - Mère	.20	-	-	-	-	-
Âge - Mère	.15	-	-	-	-	-
Total	**35.84**	**9.74**	**17.27**	**6.67**	**1.48**	**2.13**

(1) CP - Codé par la province à partir d'autres renseignements.
D - Problème de déclaration.
NR[1] - Non-réponse sur le microfilm original.
NR[3] - Total des non-réponses.

TABLEAU EXPLICATIF III. Mortinaissances - Résultats pondérés

Caractéristiques	Erreur de dépouillement		Autres incompatibilités(1)				
	Genre inconnu	Codage	M	D	NR[1]	NR[3]	IMP
	pourcentage						
Provinces utilisant des microfilms/ formules:							
Sexe	-	-	-	-	.47	-	-
Date de mortinaissance	.40	-	-	2.46	-	-	-
Genre - Accouchement	.40	1.20	-	-	-	-	-
Rang de naissance	-	-	-	-	.47	-	-
Résidence - Mère	-	1.26	.93	3.86	1.99	-	-
Durée de la grossesse	1.26	-	-	.47	-	-	-
Total - Enfants - Nés	2.80	-	.47	1.87	-	-	-
Nombre d'enfants - Nés vivants	1.40	-	.47	.47	-	-	-
Nombre d'enfants - Mort-nés	.93	-	.47	.93	-	-	-
Poids à la naissance	.47	.93	-	.47	-	3.73	-
État matrimonial - Mère	-	-	.47	-	-	-	79.15
Parents mariés l'un à l'autre	.47	-	-	-	-	-	25.12
Lieu de naissance - Père	.40	-	-	2.33	-	-	-
Date de naissance - Mère	-	-	-	.47	-	-	25.21
Âge - Mère	-	-	-	-	-	-	-
Cause du décès	-	16.39	-	.80	-	-	-
Total	**8.53**	**19.78**	**2.81**	**14.13**	**2.93**	**3.73**	**129.48**

(1) M - Modification.
D - Problème de déclaration.
NR[1] - Non-réponse sur le microfilm original.
NR[3] - Total des non-réponses.
IMP - Imputation.

TABLEAU EXPLICATIF IV. Décès - Résultats pondérés

Caractéristiques	Erreur de dépouillement		Autres incompatibilités(1)					
	Genre inconnu	Codage	M	CP	D	NR[1]	IMP	ILL
	pourcentage							
Provinces utilisant des bandes:								
Sexe	.09	-	-	-	-	.17	-	-
Lieu du décès	1.08	4.80	1.17	-	.90	1.80	-	.69
Résidence - Personne décédée	-	9.65	.30	.22	1.94	.69	-	.43
État matrimonial	.60	-	.22	-	.47	.09	-	-
Date de naissance	2.91	.22	-	.17	.17	.17	12.58	.47
Âge	1.29	.22	.39	-	.43	1.54	-	.34
Date du décès	.09	.09	-	-	.09	1.63	11.71	.82
Cause du décès	-	7.01	-	-	-	-	-	.34
Autopsie	3.69	-	-	-	-	-	-	.09
Date de certification	.22	-	-	-	-	-	-	.22
Total	**9.97**	**21.99**	**2.08**	**.39**	**4.00**	**6.09**	**24.29**	**3.40**
Provinces utilisant des microfilms/formules:								
Sexe	.25	-	-	-	-	-	-	.07
Lieu du décès	-	3.23	.25	-	2.14	.13	-	.39
Résidence - Personne décédée	-	3.16	1.50	-	.59	-	-	.13
État matrimonial	.20	-	-	-	.07	-	-	.13
Date de naissance	.52	.25	1.00	-	.32	-	1.69	.32
Âge	1.19	-	.50	-	.07	-	-	.13
Date du décès	.58	.25	.32	-	-	-	-	.20
Cause du décès	1.64	8.73	.25	-	-	-	-	.72
Autopsie	.25	-	-	-	-	-	-	.07
Date de certification	.38	-	-	-	-	-	-	.26
Total	**5.01**	**15.62**	**3.82**	**-**	**3.19**	**.13**	**1.69**	**2.42**
Toutes les provinces:								
Sexe	.10	-	-	-	-	.15	-	.01
Lieu du décès	.97	4.64	1.08	-	1.03	1.62	-	.66
Résidence - Personne décédée	-	9.00	.42	.19	1.81	.63	-	.40
État matrimonial	.56	-	.19	-	.43	.08	-	.01
Date de naissance	2.67	.22	.10	.15	.19	.15	11.48	.43
Âge	1.28	.19	.40	-	.39	1.39	-	.32
Date du décès	.14	.10	.03	-	.08	1.46	10.56	.75
Cause du décès	.17	7.18	.03	-	-	-	-	.38
Autopsie	3.35	-	-	-	-	-	-	.08
Date de certification	.23	-	-	-	-	-	-	.22
Total	**9.47**	**21.33**	**2.25**	**.34**	**3.93**	**5.48**	**22.04**	**3.26**

(1) M - Modification.
CP - Codé par la province.
D - Problème de déclaration.
NR[1] - Non-réponse sur le microfilm.
IMP - Imputation.
ILL - Illisible.

POOR

QUALITY

ORIGINAL